Anais do 6º Congresso Brasileiro de Agroinformática – SBIAgro 2007, São Pedro, Brasil, 8-11 outubro 2007, Embrapa Informática Agropecuária, p.221-225.

RECUPERAÇÃO DE INFORMAÇÃO: BASES DE DADOS DA PESQUISA AGROPECUÁRIA

ISAQUE VACARI[1]
LEILA MARIA LENK[2]
LUIS EDUARDO GONZALES[3]
MARCOS CEZAR VISOLI[4]

RESUMO: Apresentação do processo de migração do sistema de recuperação de informação das Bases de Dados da Pesquisa Agropecuária - BDPA, da arquitetura de software proprietário para software livre. O artigo relata dois casos de sucesso de aplicações na Internet para organização e disponibilização de informação (Google e Yahoo), apresenta os problemas encontrados e os requisitos para o novo sistema de busca textual, a solução tecnológica escolhida, bem como os ganhos obtidos com a migração para software livre. Apresenta, por fim, o sistema desenvolvido, com os novos recursos disponíveis e sua contribuição para a pesquisa agropecuária brasileira.

PALAVRAS-CHAVE: recuperação de informação, indexação textual, software livre, pesquisa agropecuária.

INFORMATION RETRIEVAL – EMBRAPA'S DATABASES OF AGRICULTURAL RESEARCH

ABSTRACT: Presentation of the migration process, from proprietary to free software architecture, as used by Embrapa for his information retrieval system BDPA (Databases of Agricultural Research). This poster presents two successfull applications (Yahoo and Google) of organisation and presentation of information in the web and describes problems found in the migration process, as well as the technological solution choosed for the new system, with the advantages obtained with the use free software. Finally, shows the new system and the new features developed, regarding to his contribution to the brazilian agricultural research.

KEY-WORDS: Information retrieval; textual indexing; free software; agricultural research.

1. INTRODUÇÃO

Com o aumento do uso da internet, algumas companhias, como a Yahoo (Yahoo, 2007), fizeram da organização e da disponibilização da informação, o seu negócio. A Yahoo nasceu como um catálogo ou diretório de *links*, resultado do esforço milhões de usuários da rede. A entrada da empresa Google (Google, 2007), na área de mecanismos de busca, com a técnica *PageRank,* um método de fornecer melhores resultados de busca, usando a estrutura de *links* da Internet, ao invés de somente as características dos documentos, fortaleceu o portifólio de casos de sucesso de aplicações que lidam com a organização e acesso à informação para a Internet. Não menos importante, a esse portifólio pode ser acrescentado a criação de repositórios e bibliotecas digitais.

[1] Analista, Tecnólogo em Processsamento de Dados, Faculdade de Tecnologia de Americana (FATEC-AM). E-mail: isaque@cnptia.embrapa.br
[2] Bibliotecária, MSc em Ciência da Informação, IBICT. E-mail: leila@cnptia.embrapa.br
[3] Analista, Bsc em Análise de Sistemas, PUC – Campinas, SP. e-mail: eduardo@cnptia.embrapa.br
[4] Pesquisador, BSc em Ciência da Computação, UFSC. e-mail: visoli@cnptia.embrapa.br

Nesse contexto, a Embrapa Informática Agropecuária desenvolveu o sítio de busca Bases de Dados da Pesquisa Agropecuária - BDPA (Embrapa Informática Agropecuária, 2007a) que reúne o Acervo Documental das 38 bibliotecas da Embrapa em um único sistema de recuperação de informação, abrangendo a produção científica de seus pesquisadores, as teses, a literatura adquirida pela Empresa desde sua criação, os títulos e as coleções de periódicos. Esse sistema foi concebido para expressar o conhecimento gerado e adquirido pela Embrapa e têm por objetivo contribuir para o cumprimento da missão institucional da Empresa, que é "*Viabilizar soluções para o desenvolvimento sustentável do espaço rural, com foco no agronegócio, por meio da geração, adaptação e transferência de conhecimentos e tecnologias, em benefício dos diversos segmentos da sociedade brasileira*".
O presente trabalho trata da evolução da BDPA, destacando os novos recursos de busca, performance, segurança, interface com o usuário e os benefícios advindo com o uso prioritário de tecnologias livres.

2. OBJETIVO

A BDPA vem sendo desenvolvida desde 1991 e no ano de 1996 foi disponibilizada a primeira versão para Internet. A versão 1.0 da BDPA foi construída usando arquitetura de *software* proprietária devido aos recursos de *softwares* disponíveis na época de sua concepção. Esta opção mostrou-se adequada e o sistema foi largamente utilizado. A necessidade de evoluir a plataforma tecnológica, de oferecer novos recursos e de conquistar uma maior autonomia tecnológica sobre a infra-estrutura utilizada, levaram o projeto para uma nova versão baseada totalmente em *software* livre, em acordo com o Plano Diretor da Embrapa Informática Agropecuária.

3. MATERIAL E MÉTODOS

Compõem a BDPA as bases de dados Acervo Documental Embrapa (Embrapa Informática Agropecuária, 2007b), Produção Científica Embrapa, Catálogo Coletivo de Periódicos Embrapa, Bases Temáticas Embrapa e Cadastro de Instituições que viabilizam o acesso à literatura técnico-científica existente nas bibliotecas das Unidades de Pesquisa da Embrapa, em todo território nacional. A base de dados mais significativa da BDPA é o Acervo Documental Embrapa, composta por aproximadamente 450.000 registros. A consulta ao acervo e demais bases de dados foram desenvolvidas inicialmente com o indexador proprietário Rubicon (Tamarack Associates, 2007) para o sistema operacional proprietário Windows (marca registrada de Microsoft Inc.). Devido o aumento ao acesso ao sistema de busca da BDPA, e a demanda por novos recursos de busca, e as ações para o uso de SL nas aplicações desenvolvidas pela Empresa, a arquitetura relatada apresentou deficiências como solução tecnológica para evolução da BDPA. As principais carências são descritas à seguir:

- sistema operacional proprietário, mais suscetível à invasões, e portanto mais inseguro para ser um servidor *web*;
- mecanismo indexador proprietário, com dificuldade para corrigir erros, inserir melhorias, adaptações. Além disso necessita pagamento de licenças de software;
- uso da linguagem de programação proprietária Delphi (Borland Software Corporation, 2007) para o desenvolvimento da interface do sistema de busca, que está deixando de ser uma ferramenta utilizada em projetos da Embrapa Informática Agropecuária, dado a prioridade em seu Plano Diretor de desenvolver aplicações web com ferramentas livres.

Com o conhecimento e experiência da Embrapa Informática Agropecuária no uso da plataforma J2EE (referência) e de Sistemas Gerenciadores de Banco de Dados (SGBD's) livres, o problema para migração e evolução da BDPA passou a ser a escolha do mecanismo

de indexação textual. Com base nisto foram realizados estudos sobre indexadores textuais livres e gratuitos atuando de forma integrada à SGBD's livres discutidos em Vacari & Visoli 2007a, como também, de forma independente discutidos em Vacari & Visoli 2007b. Fundamentalmente, o resultado do estudo apresenta a biblioteca Lucene (The Apache Software Foudation, 2007) como solução para indexação e busca textual em gigantescas bases de dados, principalmente para aplicações com as seguintes características:

- independência de SGBD, pois a camada de indexação é inserida entre o SGBD e a aplicação;
- suporte à todos os principais recursos de busca, discutidos em VACARI & VISOLI Visoli (2007a), como: operadores booleanos, truncagem, mascaramento, busca por proximidade, busca por frase, busca por campo específico, agrupamento, *stopwords*, indexação dos principais tipos de campo (texto, data, hora, número e visualização de dados binários), ordenação do resultado da busca por relevância e outros campos definidos pelo usuário, suporte à pesquisa acentuada, marca automática sobre o texto encontrado etc.;
- criação de um banco de dados indexado para busca, cujos dados são provenientes de documentos de texto, HTML, PDF, de banco de dados, ou de qualquer outro formato não texto que possa ser convertido em formato texto através de um filtro adequado;
- disponibilidade de várias linguagens de programação para o desenvolvimento da interface com o usuário: Java, Perl, Python, C++, .NET e Ruby;
- sem custos com licença e autonomia tecnológica, pois o Lucene é disponibilizado como SL.

Fruto dos ótimos resultados obtidos com o indexador Lucene, detalhados em VACARI & VISOLI (2007b), em termos de desempenho, performance, recursos de busca e facilidade de desenvolvimento de aplicações, decidiu-se usar o Lucene para implementar uma nova versão da BDPA com SL. Também, para garantir a segurança, estabilidade e performance do novo sistema foram realizados testes de sobrecarga. Foram simulados diversos ambientes de busca, todavia todos ambientes foram projetados para executar inúmeras buscas por minuto com diversos usuários conectados simultaneamente ao novo sistema. Ao fim da execução dos testes de sobrecarga o novo sistema da BDPA estava em condições normais de uso e o tempo necessário para efeturar as buscas foram altamente satisfatórias (buscas complexas foram executadas em menos de 0,5 segundos). Em detalhes, o principal ambiente usado para os testes de sobrecarga simulou cerca de 150 usuários (distribuídos em 10 computadores) conectados simultaneamente ao novo sistema de busca executando consultas ininterruptamente até o limite de 110.000 buscas, terminado a execução dos testes, os arquivos de *logs* foram analisados e cerca de 75% das buscas foram executadas em menos de 0,5 segundos. A incorporação dos testes de sobrecarga no processo de desenvolvimento do *software* foi de grande importância para garantir a qualidade e estabilidade do novo sistema de busca da BDPA para seus usuários.

O novo sistema da BDPA está disponível em <http://www.bdpa.cnptia.embrapa.br>, com a seguinte arquitetura tecnológica usando *softwares* gratuitos:

- sistema operacional: FreeBSD;
- linguagem de programação: Java com Java Server Pages;
- mecanismo de indexação e busca textual: Lucene;
- servidor *web*: Apache TomCat Web Server.

Figura 1. Novo sítio de busca das Bases de Dados da Pesquisa Agropecuária

4. CONCLUSÕES E SUGESTÕES

A disponibilização de dados e aplicações na Internet por instituições de pesquisa e ensino (como é o caso da Embrapa com a BDPA) tem sido uma prática bem sucedida e aceita pela comunidade de ensino e pesquisa. O desafio maior para evoluir a BDPA para SL foi encontrar ferramentas de *softwares* gratuitas para a realização das tarefas de indexação e busca textual, e também, adequadas para construção de um Sistema de Recuperação de Informação para o ambiente *web*, com facilidade para implementação de novos recursos, rapidez para realização de buscas simples e complexas, integração com linguagens de programação gratuitas (como Java) e integração com Sistemas Operacionais mais apropriados e seguros para Internet (como FreeBSD e GNU/Linux). Baseado nos requisitos acima, a ferramenta de *software* escolhida foi o Lucene, uma solução livre e gratuita para indexação e busca textual em Sistemas de Recuperação de Informação.

A experiência com ferramentas livres, como o Lucene e demais *softwares* livres utilizados nesse estudo, permitiu à Embrapa Informática Agropecuária ter o domínio tecnológico no uso de indexadores textuais. Como também a arquitetura empregada para o desenvolvimento do sistema da BDPA aliada à boas práticas de desenvolvimento de *software* (como testes de automatizados de sobrecarga, uso de CVS para controle de versão e processo de desenvolvimento aberto com a utilização do banco de projetos da Embrapa Informática Agropecuária), possibilitou a escalabilidade para a construção de novos sítios *web* de busca de informações, como Produção Científica Embrapa (Embrapa Informática Agropecuária, 2007c) e Rede de Bibliotecas da Área de Engenharia e Arquitetura (Embrapa Informática Agropecuária, 2007d) com garantia de qualidade.

5. REFERÊNCIAS BIBLIOGRÁFICAS

BORLAND SOFTWARE CORPORATION. Borland IDE. Delphi Windows .NET application development too w/object Relational Mapping. Disponível em: <http://www.borland.com/br/products/delphi/index.html>. Acesso em: 03 mai. 2007.

EMBRAPA INFORMÁTICA AGROPECUÁRIA. **Bases de Dados da Pesquisa Agropecuária**. Disponível em: <http://www.bdpa.cnptia.embrapa.br>. Acesso em: 26 abr. 2007a.

EMBRAPA INFORMÁTICA AGROPECUÁRIA. **Bases de Dados da Pesquisa Agropecuária: Acervo Documental Embrapa**. Disponível em: <http://www.bdpa.cnptia.embrapa.br/index.jsp?url=acervo.jsp&baseDados=ACERVO>. Acesso em: 26 abr. 2007b.

EMBRAPA INFORMÁTICA AGROPECUÁRIA. **Produção Científica Embrapa**. Disponível em: <http://www.prodemb.cnptia.embrapa.br>. Acesso em: 26 abr. 2007c.

EMBRAPA INFORMÁTICA AGROPECUÁRIA. **Rede de Bibliotecas da Área de Engenharia e Arquitetura**. Disponível em: <http://www.rebae.cnptia.embrapa.br>. Acesso em: 26 abr. 2007d.

GOOGLE. **Google Brasil**. Disponível em: <http://www.google.com.br>. Acesso em: 26 abr. 2007.

SEARCH TOOLS CONSULTING. **Alphabetical List of SearchTools Product Reports**. Disponívem em: <http://www.searchtools.com/tools/tools.html>. Acesso em: 23 abr. 2007.

TAMARACK ASSOCIATES. **Rubicon full text search**. Disponível em: <http://www.tamaracka.com>. Acesso em: 03 mai. 2007.

THE APACHE SOFTWARE FOUNDATION. **Lucene**. Disponível em: <http://lucene.apache.org>. Acesso em: 23 abr. 2007.

VACARI, I.; VISOLI, M. C. **Indexação Textual usando Sistemas Gerenciadores de Banco de Dados Livres**. Campinas: Embrapa Informática Agropecuária, 2005. 6 p. (Embrapa Informática Agropecuária. Comunicado Técnico, 70). Disponível em: <http://www.cnptia.embrapa.br/modules/tinycontent3/content/2005/comtec70.pdf>. Acesso em: 23 abr. 2007a.

VACARI, I.; VISOLI, M. C. **Indexação Textual no Sistema de Bases de Dados da Pesquisa Agropecuária**. Campinas: Embrapa Informática Agropecuária, 2006. 6 p. (Embrapa Informática Agropecuária. Comunicado Técnico, 71). Disponível em: <http://www.cnptia.embrapa.br/modules/tinycontent3/content/2006/ct71.pdf>. Acesso em: 23 abr. 2007b.

YAHOO! DO BRASIL INTERNET LTDA. **Yahoo! Brasil**. Disponível em: <http://br.yahoo.com>. Acesso em: 26 abr. 2007.